Report

Ahemd Elharith Osama

2024-06-28

Introduction

This statistical report is the conclusion of deep statistical analysis and hypothesis testing that comes after the six following phases

1. Discovering the data.
2. Explanatory Data Analysis “EDA”.
3. Cleaning the data.
4. Data analysis and visualization.
5. Statistical testing
6. Insights and conclusion(Report.pdf).

These phases are included in detailed manner is this document : Crop-yield-analysis.docx”.

We’ll show the most important insights and outcomes of 3 , 4 and 5.

Insights and summary

3.Cleaning the data outcomes:

1. The total number of rows 2826 and number of columns 29

4.Data analysis and visualization outcomes:

1. Numerical features

   From the pots and graphs we see that the data is very skewed .

2. categorical features
    1. From the pots and graphs we see that the data is imbalance.

4.Statistical testing outcomes:

1. What the distribution of the yield variables ?

```
+----------------+-----------------+------------------------+
| Test statistic |     P value     | Alternative hypothesis |
+================+=================+========================+
|     0.2062     | 9.166e-105 * * * |       two-sided        |
+----------------+-----------------+------------------------+

Table: Asymptotic one-sample Kolmogorov-Smirnov test: `train_df$Yield`
```

   conclusion: With a low test statistic D=0.2062D=0.2062 and an exceedingly small p-value (`9.166e-105 * * *`), there is strong evidence to reject the null hypothesis. Therefore, we conclude that the distribution of `train_df$Yield` significantly deviates from the theoretical distribution specified in the test (likely a standard normal distribution).

2. what is the crop yield for each District in India what is the largest ?

```
+----------------+----+-----------------+
| Test statistic | df |     P value     |
+================+====+=================+
|     744.1      | 3  | 5.745e-161 * * * |
+----------------+----+-----------------+

Table: Kruskal-Wallis rank sum test: `Yield` by `District`
```

   conclusion: **P Value**: The p-value assesses the probability of observing a test statistic as extreme as HH, assuming that the null hypothesis (no difference in medians) is true. A p-value of `5.745e-161 * * *` indicates an extremely small value (much less than 0.0010.001), suggesting strong evidence against the null hypothesis.

3. What is the different agriculture methods that influence the crop yield ?

```
+----------------+-----------------+------------------------+
| Test statistic |     P value     | Alternative hypothesis |
+================+=================+========================+
|     71588      | 4.443e-08 * * * |       two.sided        |
+----------------+-----------------+------------------------+

Table: Wilcoxon rank sum test with continuity correction: `Yield` by `Harv_me
thod`
```

conclusion: The p-value assesses the probability of observing a test statistic as extreme as WW, assuming that the null hypothesis (no difference in medians) is true. A p-value of `4.443e-08 * * *` indicates a very small value (much less than 0.0010.001), suggesting strong evidence against the null hypothesis.

4. What the best agriculture methods that implies better crop yield ?

```
+----------------+---------+-----------------------+
| Test statistic | P value | Alternative hypothesis |
+================+=========+=======================+
|     71588      |    1    |        greater         |
+----------------+---------+-----------------------+

Table: Wilcoxon rank sum test with continuity correction: `Yield` by `Harv_me
thod`
```

conclusion: The p-value assesses the probability of observing a test statistic as extreme as WW, assuming that the null hypothesis (no difference in medians) is true. A p-value of `1` indicates that the observed data is as likely to occur under the null hypothesis as under the alternative hypothesis.

5. What are the variables that are correlated to the yield variable ?

```
+----------------+---------+-----------------------+--------+
| Test statistic | P value | Alternative hypothesis |  rho   |
+================+=========+=======================+========+
|   283225039    | 0 * * * |       two.sided        | 0.9247 |
+----------------+---------+-----------------------+--------+

Table: Spearman's rank correlation rho: `train_df$Yield` and `train_df$Acre`
```

- conclusion : With a high test statistic and a very small p-value, there is strong evidence to reject the null hypothesis. Therefore, we conclude that `train_df$Yield` and `train_df$Acre` are significantly positively correlated in a monotonic manner